# Communication skills training exploiting multimodal emotion recognition

# Communication Skills Training Exploiting Multimodal Emotion Recognition

Kiavash Bahreini[a*], Rob Nadolski[a], and Wim Westera[a]

[a]*Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences, Open University of the Netherlands*

[*]Kiavash Bahreini, Welten Institute, Research Centre for Learning, Teaching and Technology, Faculty of Psychology and Educational Sciences, Open University of the Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands. Email: kiavash.bahreini@ou.nl.

**Abstract**

The teaching of communication skills is a labour-intensive task because of the detailed feedback that should be given to learners during their prolonged practice. This study investigates to what extent our FILTWAM facial and vocal emotion recognition software can be used for improving a serious game (the Communication Advisor) that delivers a web-based training of communication skills. A test group of 25 participants played the game wherein they were requested to mimic specific facial and vocal emotions. Half of the assignments included direct feedback and the other half included no feedback. It was investigated whether feedback on the mimicked emotions would lead to better learning. The results suggest the facial performance growth was found to be positive, particularly significant in the feedback condition. The vocal performance growth was significant in both conditions. The results are a significant indication that the automated feedback from the software improves learners' communication performances.

**Keywords**

Assessment of learners; Communication skills; Emotion recognition; Serious gaming; Feedback.

# 1 Introduction

In the last decades communication skills have become much more important in a wide variety of jobs' portfolios (e.g. Brantley & Miller, 2008). This pattern can be easily explained as a consequence from today's networked world of digital technologies, which have considerably altered the nature of professional work. Increasingly, professional labour has become knowledge-driven and requires extensive collaboration and communication between professionals. For instance, a few decades ago engineers could more or less work independently without intensive communication with fellow engineers. But, as products and processes have become increasingly complex, diverse specialists from a wide range of disciplines have to work closely together in order to develop or support such sophisticated products and services.

Nowadays, communication skills' training is mostly arranged in face-to-face contexts. But such training is not without problems. Communication courses require intensive skilled tutoring by teachers, preferably at a one-to-one personalised level, which puts heavy loads on tutoring capacity. Moreover, because of this heavy tutoring load, communication courses are expensive, or even worse; they are less effective because of a teacher bandwidth problem (Hager, Hager, & Halliday, 2006; Vorvick, Avnon, Emmett, & Robins, 2008). Another issue is the required teacher quality: providing proper feedback in communication training is a

challenging and demanding task that is only mastered after appropriate training (Cantillon & Sargeant, 2008). The shortage of well-qualified teachers is problematic (Hager et al., 2006). In addition, face-to-face training is usually arranged within fixed schedules and time slots, which severely affects the learners' flexibility to do their exercises. Prolonged practice, which is an essential condition for truly mastering communication skills, also seems to be hampered by the limitations of face-to-face training. Online approaches for communication skills training would provide a more flexible training context, allowing learners to do their exercises whenever they want and wherever they are. Digital serious games developed for learning purposes facilitate such flexibility in an attractive way. Serious games are part of the wider area of the game-based learning. Abt (1970) introduced the term 'serious games' to indicate games for job training, such as the training of army personnel or insurance salesmen. Moreover, the effectiveness of serious games for learning has been acknowledged in several studies (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Wouters, Van Nimwegen, Van Oostendorp, & Van der Spek, 2013).

The current study focuses on the empirical validation of our emotion recognition approach described in previous studies (Bahreini, Nadolski, & Westera, 2014a; Bahreini, Nadolski, & Westera, 2014b; Bahreini, Nadolski, & Westera, 2015a; Bahreini, Nadolski, & Westera, 2015b) within the practical context of a serious game for communication skills training (the Communication Advisor). Our basic research question is to what extent facial and vocal emotion recognition software can be used for improving communication skills training. The Communication Advisor allows learners to practice and improve their skills for emotion expression. The Communication Advisor has been developed based on the EMERGO game engine[1]. The Communication Advisor offers learners a number of separate assignments that require them to respond to a specific problem situation by displaying pre-defined emotions. The manifested emotions of the learners are captured with the computer's webcam and microphone and are fed into the FILTWAM face and voice emotion recognition technologies. The FILTWAM software assesses the manifested emotions for both facial expression and voice intonation. The outcome is then used for presenting direct feedback to the learners about the correctness of their performances. Importantly and in contrast with physiological sensor-based emotion recognition technologies, the FILTWAM software allows for the unobtrusive in-game assessment of learners' manifested emotions, as it only captures and analyses webcam and microphone signals. Moreover, FILTWAM evaluates the observed emotion in real-time, whereby it can be used for giving direct feedback. For this validation study we used a within-subjects experimental design with 25 participants playing the Communication Advisor. Besides collecting the players' performance data, we collected qualitative data comprising the players' appreciations and judgements about the game as a learning tool.

We first provide a brief overview of previous research in emotion recognition used in computer-based learning. After a brief explanation of the FILTWAM framework, we describe the research methodology and the research findings. We conclude the paper by discussing the findings and making suggestions for future research.

# 2 Related work

It is commonly acknowledged that emotions are a significant influential factor in the process of learning, as they affect memory and action (Pekrun, 1992). The influence of emotions on learning is traditionally well recognized in classroom teaching practice (Bower, 1981). More recently, emotions have also received attention in the domain of intelligent tutoring systems (ITS) (Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008). An ITS is a computer-based system that is capable of providing immediate and personalized instruction and feedback to learners (Psotka & Mutter, 1988). The general premise is that extending ITS with emotion recognition capabilities would lead to better conditions for learning, as it allows for adjusting its interventions to the emotional states of the user. Although there are many studies reported in the wider domains of emotion recognition and ITS, to our knowledge no study has yet been conducted that specifically combines automatic facial and vocal emotion recognition in communication skills training.

An important success factor in classroom learning is the capability of a teacher to timely recognize and respond to the affective states of their learners. For this, teachers continuously adjust their teaching behaviour by observing and evaluating the behaviour of the learners, including their facial expressions, body movements, and other signals of overt emotions. In e-learning, just as with classroom learning, it is

[1] www.emergo.cc

not only about cognition and learning, but also about the interdependency of cognition and emotion. These relationships between learners' cognition and emotion are influenced by the electronic learning environment, which mediates the communication between participants (teacher, learner, and his peers) and contains or refers to e-learning materials (e.g., text, photos, audios and videos, and animations). Contemporary, instructional approaches increasingly address emotional dimensions by accommodating challenges, excitement, ownership, and responsibility among other things in the learning environment. Software systems for e-learning (e.g., ITS, serious games, personal learning environments) could better foster learning if they also adapt the instruction and feedback to the emotional state of the learner (Sarrafzadeh, et al., 2008). Within the scope of ITS, Feidakis and his colleagues (Feidakis, Daradoumis, & Caballe, 2011) categorized emotion measurement into three types of tools, which have been described in several previous studies: 1) psychological (Wallbott, 1998), 2) physiological (Kramer, 1991), and 3) motor-behavioural (Leventhal, 1984). Psychological tools are self-reporting tools for capturing the subjective experience of emotions of users. Physiological tools comprise sensors that capture an individual's physiological responses. Motor-behaviour tools for emotion extraction use special software to measure behavioural movements captured by PC cameras, mouse or keyboard. Most of these emotion recognition tools suffer from limited reliability and unfavourable conditions of use, which hampered successful implementation of so-called affective tutoring systems (ATS). But more recently, there has been a growing body of research on ATS that recommends emotion recognition technologies based on facial expressions (Ben Ammar, Neji, Alimi, & Gouardères, 2010; Wu, Huang, & Hwang, 2015) and vocal expressions (Rodriguez, Beck, Lind, & Lok, 2008; Zhang, Hasegawa-Johnson, & Levinson, 2003).

Communication skills' training typically involves expressing specific emotions at the right point and time; such training can become tedious, as it requires prolonged practice. Serious games offer a challenging and dynamic learning context that seamlessly combine emotion and cognition (Westera, Nadolski, Hummel, & Wopereis, 2008). Such games are characterised by timely feedback to cater for skills learning from prolonged practice and are praised for their motivational affordances (Van Eck, 2010). Please note that as online communication is inherently truncated communication, which tends to strip messages from their emotional dimensions (Westera, 2013), emotion recognition is an emerging field in human-computer interaction as this would be a promising next step in enhancing the quality of online interaction and communication. Unfortunately, only a few studies address emotion recognition in digital serious games. A study by Hyunjin and his colleagues (Hyunjin, Sang-Wook, Yong-Kwi, & Jong-Hyun, 2013) investigated whether a simple brain computer interface with a few electrodes can recognize emotions in more natural settings such as playing a game. They invited 42 participants to play a brain-controlled video game wearing a headset with single electrode brain computer interface and provided a self-assessed arousal feedback at the end of each round. By analysing the data obtained from the self-evaluated questionnaires and the recordings from the brain computer interfaces device, they proposed an automatic emotion recognition method that classifies four emotions with accuracy of about 66%. Some studies address adaptation in games based on the measurement of user's emotions, motivation, and flow (Pavlas, 2010; Tijs, Brokken, & IJsselsteijn, 2009). In the study conducted by Tijs and colleagues, the researchers investigated the relations between game mechanics, a player's emotional state, and the associated emotional data. The researchers manipulated speed as a game mechanic in the experimental sessions. They requested players to provide their emotional state for valence, arousal, and boredom-frustration-enjoyment. Moreover, they measured a number of physiology-based emotional data features. Then, they compared the previous approaches and found correlations between the valence/arousal self-assessment and the emotional data features. Finally, they found that there are seven emotional data features, such as keyboard pressure and skin conductance that can distinguish between boring, frustrating, and enjoying game modes.

Other studies have shown that it is possible to measure facial and vocal emotions with considerable reliability in real time, both separately and in combination (Bahreini et al., 2014a; Bahreini et al., 2015a; Bahreini et al., 2015b). Taking the previous research into account, our approach will use common low-cost computer webcams and microphones rather than dedicated sensor systems for emotion detection. First, emotion detection could be used for tracking the learner's moods during their learning, which could inform the pedagogical intervention strategies to be applied for achieving optimal learning outcomes. Second, when emotions are part of the learning content, which is the case in communication skills training, emotion recognition could be used for measuring the learners' mastery of emotions and providing feedback. In this study, we focus on the latter usage of the facial and the vocal emotion recognition technologies. Furthermore, we will revert to the emotion classification suggested by Ekman and Friesen (1978), which is widely used in psychological research and practice. This classification comprises of six basic emotions: happiness, sadness, surprise, fear, disgust, and anger. In addition we will include the complement, neutral emotion.

# 3 The FILTWAM framework

The FILTWAM framework enables the real-time recognition of emotions, either from facial expressions or vocal intonations. The use of the FILTWAM framework in a learning situation is shown in Figure 1. It includes five layers and a number of components within the layers. The first FILTWAM layer refers to the learner, who is the subject using the computer for accessing learning materials for personal development or preparing for an exam. The device layer reflects the equipment of the learner, whether a personal computer, a laptop, or a smart device. It is supposed to include a webcam and a microphone for collecting user data. The web interface runs a serious game (or any other online training) and allows the learner to interact with the game components. The learner will receive the feedback/content through the Internet. The web service client in the device layer uses the affective computing tool and calls the web service in the application layer. It reads the affective data and broadcasts the live stream including the facial emotion recognition expression and the vocal emotion recognition of the learner through the Internet to the web service. The affective computing tool processes facial expressions and vocal intonations data of the learner. The component 'emotion recognition from facial features' extracts facial features from the face and classifies emotions. It leads to the recognition and categorization of a specific emotion. The process of emotion recognition from facial features starts at the face detection. Then the facial feature extraction extracts a sufficient set of feature points of the learner. Finally, the facial emotion classification analyses video sequences and extracts an image of each frame for its analysis and compares the image with the data set. Its development is based on the FaceTracker software (Saragih, Lucey, & Cohn, 2011). It supports the classification of six basic emotions (Ekman & Friesen, 1978) plus the neutral emotion, but can in principle also recognize other or more detailed face expressions when required. The data layer physically stores the facial and the vocal corpuses of the emotions. The network layer uses the Internet to broadcast a live stream (cut in pieces, digitised, and sent) of the learner. The application layer consists of an e-learning environment and its two sub components. In our case the e-learning environment was a game. The e-learning environment uses the live stream of the facial and the vocal data of the learner to deliver new content. The web service receives emotional data from the web service client. The rules engine component in the game manages didactical rules and triggers the relevant rules for providing feedback as well as tuned training content to the learner via the device. This component uses some decision algorithms to provide feedback. At this stage, learners can receive a feedback based on their facial and vocal emotion expressions. For voice data the process is largely similar. A detailed description for voice emotion recognition is available in a previous study (Bahreini et al., 2015a).

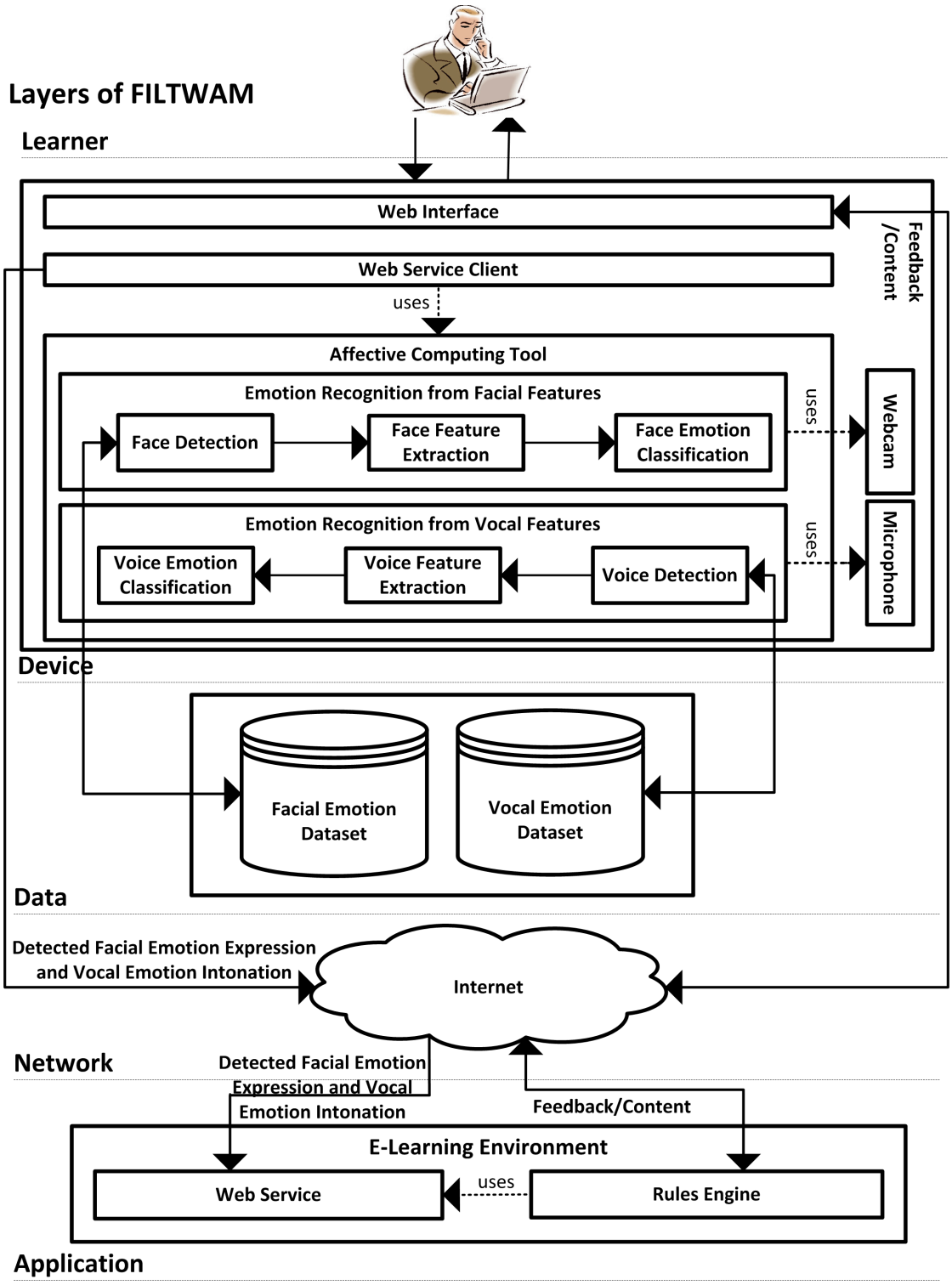


***Figure 1.*** *The FILTWAM framework for multimodal emotion recognition in an e-learning environment.*

# 4 Methodology

The Communication Advisor is the web-based serious game that was used in our study. All participants were offered three rounds of the same 28 assignments within the game. The assignments were presented in the same order in each round. Each assignment required the participant to respond in a predefined way to a specific problem situation. This situation was briefly represented in a video clip. Participants' manifested emotions (i.e., performance as expressed facial or vocal emotion) in response to the 28 assignments that were recorded by the emotion recognition software and assessed as being correct or incorrect.

## 4.1 Participants

Twenty-five participants, all employees from the Welten Institute (16 male, 9 female; mean age = 44, standard deviation = 15) volunteered to participate in the study. Participants were non-actors. The participants were invited to do a communication skills training that constituted of a completion of all assignments within the Communication Advisor. By signing an agreement form, the participants allowed us to record all their data that could be gathered by the game, including their facial expressions and their vocal intonations. For participating in this experiment, no specific background knowledge was required.

## 4.2 Materials

The Communication Advisor is a web-based serious game for learning. It 1) deals with authentic real life tasks (challenge and real world relevance), 2) uses video to establish a real life setting (context that enables transfer of learning), 3) *can* provide immediate and frequent feedback (guides learning), 4) offers a score mechanism (quantifiable outcomes that guide learning), and 5) presents many small assignments (challenges) that require user input (interactivity).

The Communication Advisor includes several components and a number of buttons in the GUI level. Figure 2 represents the main components (Indicator, Text description, Video, Task, and Instruction) and the two buttons (Refresh and Microphone) of the Communication Advisor. Figure 3, represents the main components (Indicator, Text description, Video, Task, and Instruction), the feedback component, and the navigation button of the Communication Advisor. The three components on the top of Figure 2 are the indicator components. They simply indicate the round number, the group number, the task number within the group, and the task number of the whole tasks within each round. Every single assignment in the game included a text description of a real life situation. These texts are included in the text description component. The video component plays a video of a communication partner related to the specific task, while the task component displays a sentence with a requested emotion that is to be expressed by the learner. The instruction component shows the learner how to proceed in the game. The refresh button will replay the video of the communication partner again. The microphone button will allow the learner to record his facial emotion expressions and his vocal emotion intonations. The feedback component gives the learner his emotional feedback based on his facial expressions and his vocal intonations. The navigation button will forward the learner to the next task. After the third round is finished, the Communication Advisor provides an overview of the learners´ performances on the rounds and tasks.

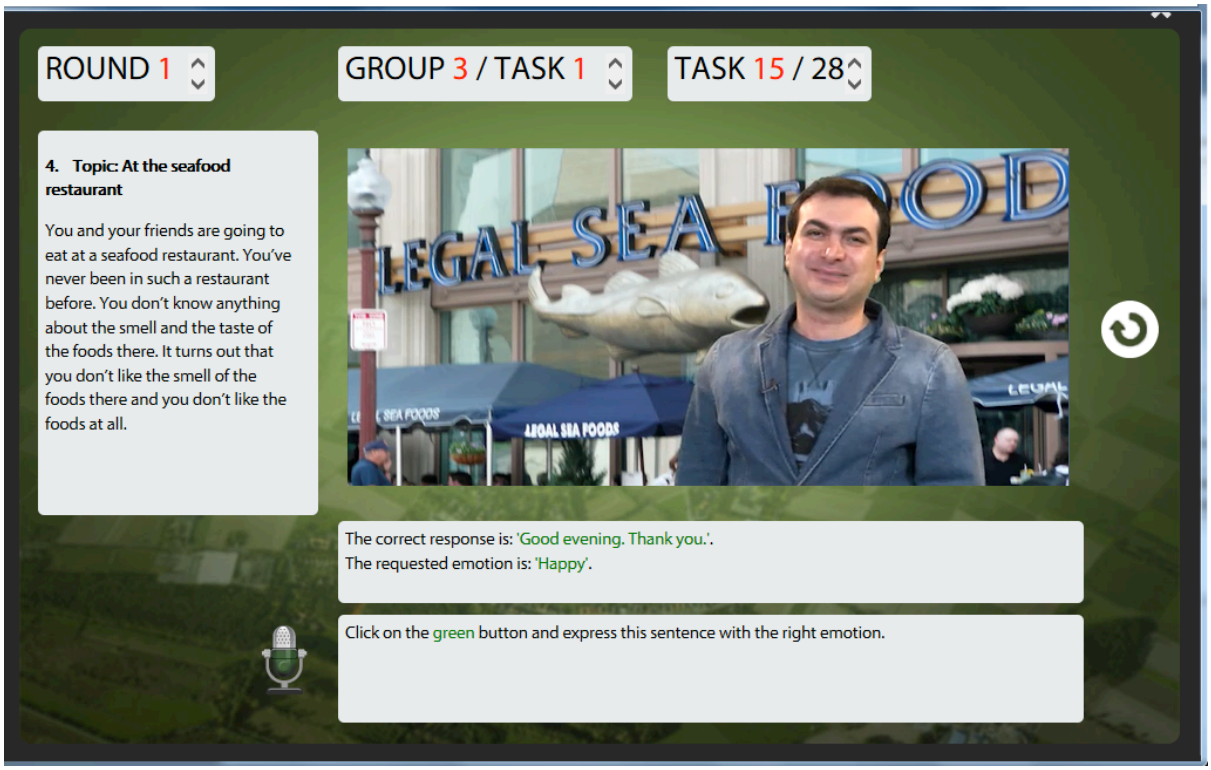


***Figure 2.*** *The game components and buttons in an assignment in the Communication Advisor.*


**Figure 3.** The game components and the navigation button in an assignment with direct feedback on vocal and facial emotions of a participant in the Communication Advisor.

The assignments in the game aimed at helping the participants to understand and improve their facial expressions and vocal intonations. After the first round, the same assignments were then again presented in round two and round three (cf. Figure 4). Every single assignment included a text description of a real life situation, a video of a communication partner (see Figure 2), and a sentence with a requested emotion that was to be expressed. The participants were asked to mimic the facial expressions while looking at the communication partner, speak aloud and produce the required voice emotions. After each video, the participant was asked to deliver the predefined response, which was then captured and analysed by the FILTWAM software. The detected correctness (or incorrectness) of the expressed emotion was then fed back directly to the game and presented on the screen. Each assignment addressed one out of seven basic facial and vocal expressions (happy, sad, surprise, fear, disgust, anger (Ekman & Friesen, 1978), and neutral). The assignments covered diverse themes: an employment situation, a visit to the dentist, a visit to a restaurant and a traffic accident, respectively. For instance, one of the assignments deals with a bad news conversation in the employment situation: your manager tells you that you will not be promoted to a managerial position. We used transcripts and instructions for the good-news and bad-news conversations from an existing OUNL training course (Lang & Van der Molen, 2008) and a communication book (Van der Molen & Gramsbergen-Hoogland, 2005).

## 4.3 Design

The experiment was arranged in a within-subjects (repeated measures) design, with two experimental conditions: 1) assignments without feedback, 2) assignments with feedback. First, in a within-subjects design, all participants are exposed to all conditions, which means that differences between conditions are not blurred by individual differences in e.g. personality, acting skills, or emotional intelligence. Second, for practical reasons the number of participants had to be restricted; therefore it was not feasible to arrange two or more trial groups that would allow for a between groups comparison. The Communication Advisor included three consecutive rounds; each round (28 assignments) included four blocks of seven assignments. In each of the blocks all seven basic emotions were addressed. Figure 4 displays the structure.

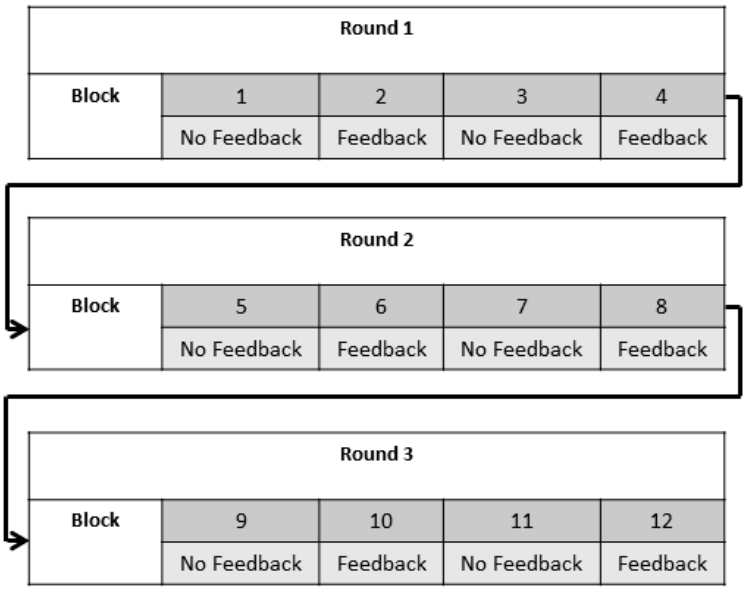


**Figure 4.** Assignments in the Communication Advisor: 12 blocks of 7 assignments.

Given this structure, in total eighty-four face expressions and eighty-four voice intonations of each participant were gathered. The two experimental conditions are alternated across the blocks of seven assignments: the assignments in the odd blocks are without feedback; the assignments in the even blocks are with feedback. Given the similarities across the assignments we assume that the difficulty levels of the assignments, or rather of the blocks, are comparable, as to allow for a comparison between experimental conditions.

## 4.4 Procedure

Participants individually performed all assignments in a single session of about 120 minutes. There were short breaks between each round for avoiding fatigue. The sessions were conducted in a silent room with

good lighting conditions. During the sessions a moderator was present in the room. The moderator gave a short instruction at the beginning of the session and asked the participants to mimic the seven basic emotions to calibrate the face emotion recognition software. For voice emotion recognition no such calibration was needed. The instruction included the request to show real expressions that are moderate and not too intensive. The performance was checked by face and voice emotion recognition software and assessed as being correct or incorrect. In the even blocks, the participants received direct feedback on the screen about the correctness or incorrectness of their expressed emotions (cf. Figure 3). After the sessions, the participants were asked to fill out an online questionnaire about their opinions and appreciations. Finally, the participants were requested not to talk to each other about the experiment in between sessions so that they would not influence each other.

## 4.5 Test environment

All assignments in the game were performed on a single web browser on a standard Windows 7 PC with integrated webcam and microphone. The EMERGO game engine was performed on a Windows server machine. The emotion recognition software applications were performed on a Mac OS computer. In principle, the experiment can also be carried out on a single computer. The data communications between the computers were performed through the web services described in the FILTWAM framework section. An external 1080HD camera was used for recording the facial and the vocal expressions of the participants and their interactions with the screen. Such data were needed for post-processing by human raters (see below), in order to be able to assess the accuracy of the emotion recognition software during the sessions.

## 4.6 Measurement instruments

The participants' performance levels of face and voice emotions were calculated both at block level and round level as a percentage of correct performances. In addition, we have developed an online questionnaire to collect participants' opinions and appreciations about the training sessions, their performances, and the feedback they received. All participants' data were collected using a 7-point Likert scale format (1=completely disagree, 7=completely agree). Participants' opinions about their assignments were gathered for: 1) difficulty to mimic the requested emotions, 2) quality of the given feedback, 3) self-confidence for being able to mimic the requested emotions, 4) clarity of the instructions, 5) attractiveness of the assignments, 6) relevancy of the assignments, 7) the graphical user interface of the game, 8) their concentration on the given assignments, 9) their acting skills, 10) their comfortableness after receiving the feedback on their performance, 11) their preference to receive feedback on their performance by a real person instead of by a computer, 12) their trust on the judgment of a real person for giving feedback more than the judgment of the computer, and 13) usefulness of the assignments to improve their communication skills. Furthermore, two questions were asked the participants to report their opinion in the descriptive format: 14) their suggestions to use such training in a real usage context and 15) their suggestions to improve the training.

## 4.7 Human raters

For being able to assess the accuracy of the emotion recognition software, two expert raters individually rated the facial and the vocal emotions of the participants' in the recorded video and audio files. Both raters have an academic level psychology background in emotion detection/recognition. Both raters are familiar and skilled with face, voice, and speech analysis. The same procedure as in our previous studies was followed to determine the accuracy of the emotion recognition system by the raters. Hence, the raters were asked to categorise and rate the recorded video and audio files of the participants for facial expressions and for vocal intonations. For supporting the rating process, the raters used the ELAN tool[2], which is a professional tool for making complex annotations on video and audio resources.

First, the raters received an instruction package for doing the ratings of the emotions based on recorded video and audio. Secondly, both raters participated in a training session where ratings of the participant were discussed to identify possibly issues with the rating task and to improve common understanding of the rating categories. Thirdly, raters assigned their individual ratings of participants' emotions for the complete set of recorded files. Fourthly, they participated in a negotiation session where all ratings were discussed to check whether negotiation about dissimilar ratings could lead to similar ratings or to sustained disagreement. Finally, the final ratings resulting from this negotiation session were contrasted with the software results for the further analysis by the main researcher. The data that the raters assigned during the initial training session were also included in the final analysis. The raters received: 1) a user manual, 2) twenty-five video and audio files of all the participants, 3) an instruction guide on how to use ELAN, and 4)

[2] https://tla.mpi.nl/tools/tla-tools/elan/

an excel file with twenty-five data sheets; each of which corresponded with one participant. The raters rated the facial expressions and the vocal intonations of the participants in the form of categorical labels covering the six basic emotions (happiness, sadness, surprise, fear, disgust, and anger) suggested by Ekman and Friesen (Ekman & Friesen 1978), as well as the neutral emotion.

# 5 Results and findings

First, we will explain the system's reliability by contrasting the FILTWAM software output and the raters' judgements of the participants' performances. Then, we will present the performances of the participants in the three rounds and analyse the differences between the two conditions (feedback, no-feedback). Finally, we will present the outcomes of the questionnaire.

## 5.1 Raters' results versus FILTWAM software

### 5.1.1 Face

The interrater reliability of the human raters was found to yield kappa = 0.894 ($p<0.001$). Therefore an almost perfect agreement among human raters was obtained, a qualification that holds for kappa values larger than 0.8 (Landis & Koch, 1977). The overall accuracy turned out to be 90%, which confirms the high quality of raters, as from the literature we know that the accuracy of human emotion recognition is around 80% (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). Furthermore, we contrasted the face software output and the human ratings using the raters' agreement about the displayed emotions as a reference. Most agreement between FILTWAM and the raters is obtained for the emotion category of happiness (kappa = 0.855, $p < 0.001$) followed by neutral 0.805, anger 0.805, disgust 0.783, sadness 0.744, surprise 0.727 and fear 0.623, respectively. The overall interrater reliability between the software and the human raters was kappa = 0.776 ($p < 0.001$). According to Landis and Koch (1977) this reflects substantial agreement, because kappa is between 0.6 and 0.8. The overall accuracy of the face software was 70%.

### 5.1.2 Voice

The interrater reliability of the two human raters was kappa = 0.881 ($p < 0.001$), which is an almost perfect agreement (Landis & Koch, 1977). The overall accuracy of their voice ratings was 89%. By contrasting the voice software output and the human ratings we found that most agreement is obtained for the emotion category of anger (Kappa = 0.856, $p < 0.001$) followed by happiness 0.813, neutral 0.740, sadness 0.731, surprise 0.658, disgust 0.585, and fear 0.575, respectively. The overall interrater reliability between the software and the human raters was kappa = 0.727 ($p < 0.001$), which is to be qualified as substantial agreement (Landis & Koch, 1977). The overall accuracy of the voice software was 61.5%.

## 5.2 Performances

### 5.2.1 Facial expression performance

Facial performances in the three consecutive rounds are presented in Table 1.

**Table 1.** Means, standard deviations, and standard errors of the face performances of the participants at each round.

| Face Performance | Mean | Std. Deviation | Std. Error |
|---|---|---|---|
| **Round 1** | 0.63 | 0.09 | 0.02 |
| **Round 2** | 0.64 | 0.12 | 0.02 |
| **Round 3** | 0.70 | 0.15 | 0.03 |

A repeated measures ANOVA shows that the face performances are significantly different across rounds, $F(2, 48)=6.48$, $p<0.025$. The data in Table 1 show a gradual increase of performance. Pairwise comparisons between 3 rounds with Bonferroni corrections show that there are significant differences between round 2 and 3 ($p=0.04$), and between round 1 and round 3 ($p=0.007$). The required assumption of sphericity of the data was established by a Mauchly's test, $x^2(2)=0.073$, $p=0.964$.

Overall, the performance of the participants showed a significant and steady increase of 11% for the facial emotion expression.

### 5.2.2 Vocal expression performance

We have followed the same procedure for voice as we have done for face. Table 2 shows the voice performances of the participants at each round.

**Table 2.** Means, standard deviations, and standard errors of the voice performances of the participants at each round.

| Voice Performance | Mean | Std. Deviation | Std. Error |
|---|---|---|---|
| **Round 1** | 0.47 | 0.10 | 0.02 |
| **Round 2** | 0.55 | 0.13 | 0.03 |
| **Round 3** | 0.63 | 0.14 | 0.03 |

The data in Table 2 show a gradual increase of voice performance. The repeated measures ANOVA shows that the voice performances are significantly different across rounds, $F(1.8, 42)=20.6$, $p<0.025$. Pairwise comparisons of the 3 rounds with Bonferroni corrections show that differences between all rounds are significant: between round 1 and 2 ($p=0.007$), between round 2 and 3 ($p=0.002$), and between round 1 and round 3 ($p=0.001$). The required assumption of sphericity of the data was established by a Mauchly's test, $x^2(2)=5.562$, $p=0.062$. The entire performance of the participants showed a significant and steady increase of 33% for the vocal emotion expression.

### 5.2.3 Facial feedback versus no-feedback

We aggregated the even blocks in each round to calculate the average performance in the feedback condition per round. Likewise, the odd blocks were used to calculate the average performance in each round in the no-feedback condition. Figure 5 represents the average facial performances in the three rounds, the no-feedback condition and the feedback condition, respectively.

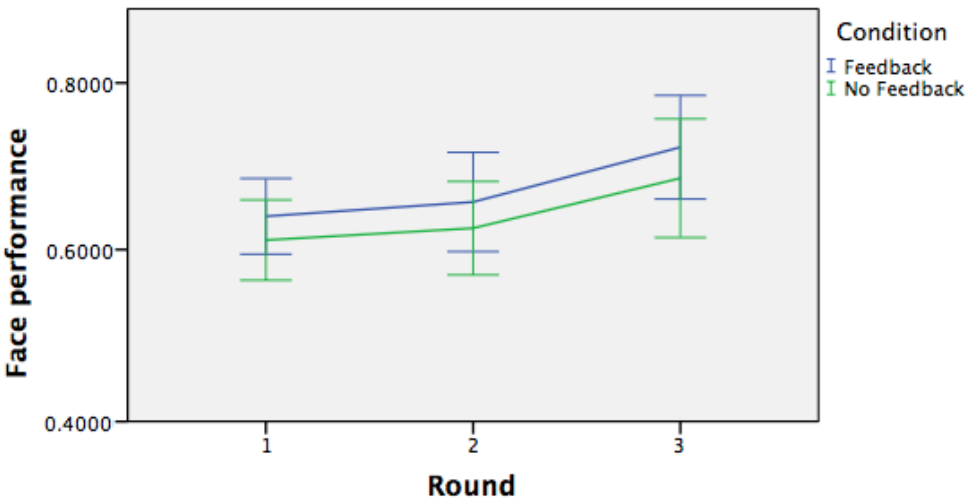


**Figure 5.** Face growth between the no-feedback and the feedback conditions.

Overall the figure suggests a better performance in the feedback condition as compared to the no-feedback condition. However, the error bars are substantial. Paired-samples t-tests between the two conditions for each round did not reveal significant differences. Likewise, the overall performances in three combined rounds did not show significant differences between the no-feedback and the feedback data.

### 5.2.4 Vocal feedback versus no-feedback

Paired-samples t-tests, between the no-feedback and feedback conditions in each round, show that vocal performances differ significantly (see Table 3). The data show a significantly higher vocal performance in the feedback condition. The data also suggest a faster growth of performance in the feedback condition.

**Table 3.** Paired-samples t-tests between the no-feedback and feedback conditions for voice performance.

| | **No-feedback Vocal Performance (SE)** | **Feedback Vocal Performance (SE)** | **T-statistic** |
|---|---|---|---|
| **Round 1** | 0.45 (0.03) | 0.50 (0.02) | t(24)=-2.22, p=0.036, r=0.41 |
| **Round 2** | 0.51 (0.03) | 0.59 (0.03) | t(24)=-2.94, p=0.007, r=0.51 |
| **Round 3** | 0.58 (0.04) | 0.67 (0.03) | t(24)=-2.80, p=0.01, r=0.50 |
| **Overall** | 0.51 (0.03) | 0.59 (0.02) | t(24)=-4.04, p=0.001, r=0.64 |

### 5.2.5 Facial performance growth across rounds

Repeated measures ANOVA for the no-feedback condition between rounds did not show significant differences, that is, the performance growth in the no-feedback condition cannot be confirmed (see Table 4).

**Table 4.** Means and standard errors of the face performances of the participants at each round for no feedback condition.

| **Face Performance** | **Mean** | **Std. Error** |
|---|---|---|
| **Round 1** | 0.61 | 0.02 |
| **Round 2** | 0.62 | 0.03 |
| **Round 3** | 0.68 | 0.03 |

In the feedback condition, however, we found a significant result (F(5.4), P=0.005). Paired-samples t-test showed that the observed facial performances in round 1 and round 3 are significantly different, growing from 0.64 to 0.72 (p=0.02).

**Table 5.** Means and standard errors of the face performances of the participants at each round for feedback condition.

| **Face Performance** | **Mean** | **Std. Error** |
|---|---|---|
| **Round 1** | 0.64 | 0.02 |
| **Round 2** | 0.66 | 0.03 |
| **Round 3** | 0.72 | 0.03 |

### 5.2.6 Vocal performance growth across rounds

A comparison of vocal performances between rounds shows significant results for both conditions. In the no-feedback condition, we found a significantly statistical performance growth between all rounds (See Table 6). In round 2, performance goes up from 0.45 to 0.51 (p=0.04); in round 3 performance rises to 0.58 (p=0.01).

**Table 6.** Means and standard errors of the voice performances of the participants at each round for no-feedback condition.

| **Voice Performance** | **Mean** | **Std. Error** |
|---|---|---|
| **Round 1** | 0.45 | 0.03 |
| **Round 2** | 0.51 | 0.03 |
| **Round 3** | 0.58 | 0.04 |

For the feedback condition we also found significant differences between all rounds. Table 7 shows the vocal performance growth in round 2 from 0.50 to 0.60 (p=0.002), and in round 3 from 0.60 to 0.67 (p=0.007).

**Table 7.** Means and standard errors of the voice performances of the participants at each round for feedback condition.

| Voice Performance | Mean | Std. Error |
|---|---|---|
| **Round 1** | 0.50 | 0.02 |
| **Round 2** | 0.60 | 0.03 |
| **Round 3** | 0.67 | 0.03 |

It should be noted that, although Bonferroni corrections have been applied throughout the analysis, the repeated use of the dataset in sections 5.2.1/3/5 and 5.2.2/4/6, respectively, would require additional corrections for the significance threshold (0.017 rather than 0.05). Although few of the results are then disqualified, the overall trends and conclusions persist.

## 5.3 Post-practice questionnaire

We follow Norman's (2010) approach to allow parametric statistics for the ordinal Likert scale data. This approach quantifies Likert scale scores, be it conditional to normality checks, and allows the scores to be represented with the arithmetical mean and standard deviation, respectively. We transformed our 7-point Likert data into a linear metric at the interval [0.0, 1.0], with the value of 0.5 as the reference of a neutral response.

First, participants did not consider themselves as particularly good actors (mean=0.39; standard deviation=0.30). With respect to the experimental conditions, they were very positive about the quality of the instructions (mean=0.88; standard deviation=0.22), the user interface (mean=0.90; standard deviation=0.20), and the arrangement of the experiment (mean=0.78; standard deviation=0.18). With respect to the communication training, participants were moderately positive about the attractiveness of assignments (mean=0.66; standard deviation=0.20), the appropriateness of the assignments for training the communication skills (mean=0.61; standard deviation=0.17), the relevancy of contents (mean=0.64; standard deviation=0.21), the confidence to mimic the requested emotions (mean=0.60; standard deviation=0.22), the helpfulness of the feedback (mean=0.67; standard deviation=0.23), and the comfortability while receiving feedback (mean=0.69; standard deviation=0.27). They reported that they were neutral about the ease of mimicking requested emotions (mean=0.53; standard deviation=0.23), trusting human judgments better or worse than judgements by the computer (mean=0.53; standard deviation=0.25), and receiving feedback from a real person or a computer (mean=0.56; standard deviation=0.22).

# 6 Discussion

In a within-subjects experiment, it was investigated whether feedback on the mimicked emotions would lead to better learning. Facial performance growth during the game was found to be positive, particularly significant in the feedback condition. The vocal performance growth was significant in both conditions, while the growth is stronger when feedback is provided. The performance of the participants showed a significant and steady increase of 11% for facial emotion expression and 33% for vocal emotion expression. This establishes the game's role as a tool for learning. These results suggest a positive contribution of the automated feedback to the process of mastery.

A principal requirement for the successful use of emotion recognition software is its accuracy. While using the judgement of human raters as a reference, the accuracies of our emotion recognition software turned out to be 70% for facial emotions and 61.5% for vocal emotions. Although these accuracies are lower than those of the human raters involved (90% for facial emotions and 89% for vocal emotions), the result is consistent with previous studies (Busso, Deng, & Yildirim, 2004; Jaimes, & Sebe, 2007; Bahreini et al., 2014a; Bahreini et al., 2015a). One may wonder if the FILTWAM emotion recognition software will work better for some emotions than for other ones. This is actually a follow up research question that is beyond the scope of this study.

For investigating the impact of automated feedback on performance we compared the performances in the feedback condition with those in the no-feedback condition in each of the rounds. For facial emotions the performances were found to be systematically higher in the feedback condition, but the differences with the no-feedback condition were not statistically significant. A similar pattern was found for vocal emotions, be it that the differences were statistically significant. The vocal performances of participants were found to be significantly higher in the feedback condition as compared to the no-feedback condition. These results suggest a positive contribution of the automated feedback to the process of mastery.
In both conditions the facial performance growth across rounds was found to be positive, be it only significant in the feedback condition. The vocal performance growths across rounds are significant in both conditions: also the growth is stronger when feedback is provided. Again this is a significant indication of the effectiveness of automated feedback.
As can be concluded from the above, both feedbacks on facial emotion expression and on vocal emotion expression result in increased performances. Yet, in some cases, particularly regarding facial emotions, we were not able to demonstrate significant effects. Although the total number of observations was high (2100 facial and 2100 vocal), the sample size of participants was only 25. A larger sample could have demonstrated more significant impact, but this was not feasible within the context of this study.
Although the 28 assignments share the same set-up, a block-wise calculation of performance scores (each round comprises 4 blocks) shows a jagged pattern rather than a gradual increase (cf. Figure 6). This may be an indicator of unequal complexity of the assignments, which would affect the validity of the study, while favouring one condition above the other.

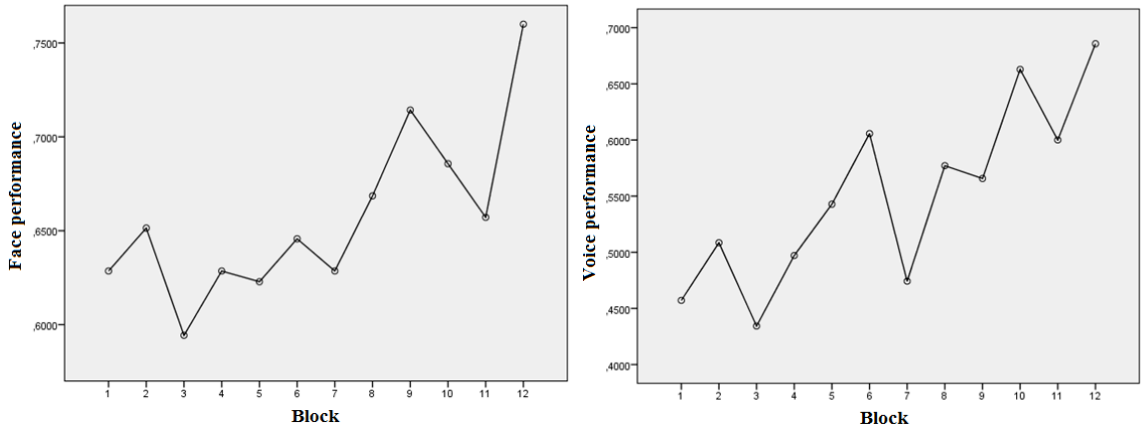


**Figure 6.** Facial and vocal performances at block level.

In a separate analysis we have cancelled out the effects of potential complexity differences of the assignments by using the 4 blocks of round 1 as a reference, that is, the performance in the blocks of round 2 and round 3 are presented as a measure relative to performances of the corresponding blocks in round 1. Figure 7 shows the relative performances of both face emotions and voice emotions.

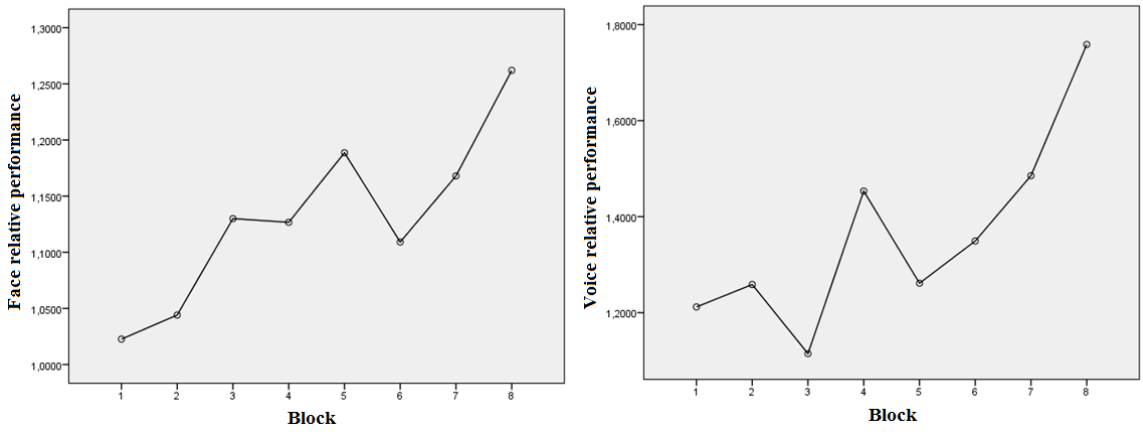


**Figure 7.** The relative performances of both face emotions and voice emotions.

Although in both graphs of Figure 7 the curves are slightly smoothened, the variability remains considerable. Statistical comparison of the no-feedback and the feedback conditions in each round did not produce significant results. Hence, correcting for complexity differences of assignments does not improve the outcomes.
As a general trend the data suggest that vocal emotion performance is structurally better that facial emotion performance. As a first explanation this may be accounted to the better accuracy of the vocal emotion recognition software: indeed, receiving correct feedback may be expected to help improve performance. Another explanation may be in the intrinsic textual setup of the assignments: reading out aloud pre-defined texts is a quite artificial task, which leaves little room for free (facial) expression, while in contrast the vocal expression is less restricted and would allow for more spontaneous performances. So far, these explanations are suppositions that could not be verified.

The results of the questionnaire indicate that the participants were ready to use a serious game to improve their communication skills instead of taking lessons from a human teacher. The results indicate that the participant were moderately positive about the helpfulness of the feedback. The participants' scores for attractiveness, appropriateness, and contents of the assignments indicate room for improvement of the Communication Advisor.

The findings obtained from this study could affect educational practice in various respects. First, it enables reliable real-time emotion detection and adaptation in e-learning. Second, it can cope with the fact that teachers are not always available to monitor progress of learners towards mastery, for example in case of communication skills training. Third, the findings seem to indicate that prolonged practice without loss of motivation is possible. Finally, the findings indicate that it is possible to integrate the Communication Advisor with ITS to achieve an ATS-based serious game environment.

In addition, the post-practice questionnaire indicates some limitations of our study that imply challenges for future research. Further improvements are possible in the area of technology development, game development, more reliable systems, and more accurate software applications. The accuracy of emotion recognition can be further improved by combining other sensory data to the FILTWAM framework. Further improvements are required to extend the FILTWAM framework for more reliable and mature exploitation of real-time emotion recognition technologies in e learning. This would offer an innovative approach for applying emotion recognition in affective e-learning (Bahreini et al., 2014a; Sebe, 2009). New software applications and serious games with emotion recognition technology could strongly influence e-learning and gaming. The current feedback mechanism in the Communication Advisor is sufficient, but restricted; therefore an improved feedback mechanism is suggested. The feedback mechanism should offer a solution, by which the learner can learn how to correct his/her mistaken facial and vocal emotional expressions. The feedback mechanism should preferably include or refer to e-learning materials (e.g., text, photos, audios and videos, and animations).

Furthermore, more ecologically valid circumstances are needed to test similar systems like the Communication Advisor with a sufficient number of participants so as to achieve full-fledged serious games. Besides the aforementioned technical challenges, legal issues, ethical issues, and essentially psychological issues also pose substantial challenges. We know that the relationship between emotion and learning is a highly complex relationship when it concerns human learning (Bower, 1981; Schwarz, 1990). The full integration and exploitation of emotion recognition techniques in e-learning environments deserves extensive investigation of that relationship.

# 7 Conclusion

The research presented in this study has shown that facial and vocal emotion recognition software can be successfully used in a serious game for communication skills training. The results indicate that, when learners repeatedly receive feedback on their assignments, their performances will improve possibly faster (as in the case of voice emotion) than when no feedback is given.
The FILTWAM emotion recognition technology connected to the Communication Advisor allowed providing real-time feedback on the learners' facial and vocal emotion expression performances. Although FILTWAM was using domestic devices (standard webcam and microphone), its accuracy was sufficient for guiding learners to better perform. Herewith the study provides a proof of concept that would allow for a wider use of the approach. Although we have considered only seven basic emotions in this study, the

FILTWAM framework can be easily extended to include more detailed emotion categories. In principle, the successful validation of FILTWAM paves the way for a structural inclusion of affective computing technologies in electronic learning environments.

# 8 Acknowledgements

We thank our colleagues at the Welten Institute of the Open University Netherlands who participated in this study. We likewise thank the two raters who helped us to rate the recorded video files. We are grateful to Hub Kurvers for designing the Communication Advisor serious game in the EMERGO environment, to Charlotte Wolff for revision of the game contents, to Jeroen Berkhout for producing the image contents and recording the video files, to Marcel Vos and Mat Heinen for technical support for video recording, to Jason Saragih and his colleagues for permission to develop the facial emotion recognition software application based on his FaceTracker software (Saragih, Lucey, & Cohn; 2011), to Setareh Habibzadeh for data entry and data visualization, and to Mieke Haemers for proofreading the manuscript. We finally thank the Netherlands Laboratory for Lifelong Learning (NELLL) of the Open University Netherlands that has sponsored this research.

# References

Abt, C. (1970). Serious games. New York: Viking Press.

Bahreini, K., Nadolski, R., & Westera, W. (2014a). Towards Multimodal Emotion Recognition in E-learning Environments. Interactive Learning Environments, 1-16.
doi= http://dx.doi.org/10.1080/10494820.2014.908927.

Bahreini, K., Nadolski, R., & Westera, W. (2014b). Multimodal Emotion Recognition for Assessment of Learning in a Game-Based Communication Skills Training. For and In Serious Games, Joint Workshop of the GALA Network of Excellence and the LEA's BOX Project at EC-TEL 2014. 22-25, September 16-19, Graz, Austria, 2014.

Bahreini, K., Nadolski, R., & Westera, W. (2015a). Towards Real-Time Speech Emotion Recognition for Affective E-Learning. Education and Information Technologies, 1-20. Springer US. doi=10.1007/s10639-015-9388-2.

Bahreini, K., Nadolski, R., & Westera, W. (2015b). Improved Multimodal Emotion Recognition for Better Game-Based Learning. A. De Gloria (Ed.): GALA 2014, LNCS 9221, 1–14. Springer International Publishing Switzerland. doi: 10.1007/978-3-319-22960-7_11.

Ben Ammar, M., Neji, M., Alimi, A. M., & Gouardères, G. (2010). The Affective Tutoring System. Expert Systems with Applications, 37(4), 3013-3023. doi: 10.1016/j.eswa.2009.09.031.

Bower, G. H. (1981). Mood and Memory. American Psychologist, 36, 129-148.

Brantley C. P., & Miller M. G. (2008). Effective Communication for Colleges. Thomson Higher Education.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech. In Proceedings of the Inter Speech, 1517-1520. Lissabon, Portugal.

Busso, C., Deng, Z., & Yildirim, S. (2004). Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, in Proceedings of ACM 6th International Conference on Multimodal Interfaces.

Cantillon, P., & Sargeant, J. (2008). Giving Feedback in Clinical Settings. BMJ, 337(a1961). doi: 10.1136/bmj.a1961.

Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T., Boyle, J.M. (2012). A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games. Computers & Education, 59 (2), 661-686. doi: 10.1016/j.compedu.2012.03.004.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: Investigator's guide. Douglas, AZ: A Human Face.

Feidakis, M., Daradoumis, T., & Caballe, S. (2011). Emotion Measurement in Intelligent Tutoring Systems: What When and How to Measure. Third International Conference on Intelligent Networking and Collaborative Systems, 807-812. Fukuoka, Japan.

Hager, P. J., Hager, P., & Halliday, J. (2006). Recovering Informal Learning: Wisdom Judgment and Community. Lifelong Learning Book Series. Springer, Dordrecht.

Hyunjin, Y., Sang-Wook, P., Yong-Kwi, L., & Jong-Hyun, J. (Oct, 2013). Emotion Recognition of Serious Game Players Using a Simple Brain Computer Interface. International Conference on ICT Convergence (ICTC), 783-786. doi: 10.1109/ICTC.2013.6675478.

Jaimes, A., & Sebe, N. (2007). Multimodal Human–Computer Interaction: A Survey, Computer Vision and Image Understanding. Special Issue on Vision for Human-Computer Interaction, 108(1-2), 116-134.

Kramer, A. F., (1991). Physiological Metrics of Mental Workload: A Review of Recent Progress. In D. L. Damos (Ed.). Multiple-Task-Performance, 329-360. London. Taylor & Francis.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33, 159-174.

Lang, G., & Van der Molen, H. T. (2008). Psychologische gespreksvoering book. Heerlen: Open University of the Netherlands.

Leventhal, H. (1984). A Perceptual Motor theory of Emotion. K.R. Scherer, P. Ekman (Eds.), Approaches to Emotion, Lawrence Erlbaum Associates, 271-291. Hillsdale, NJ.

Norman, G. (2010). Likert Scales, Levels of Measurement and the ‘‘Laws’’ of Statistics. Advances in Health and Science Education, 15(5), 625-632. doi 10.1007/s10459-010-9222-y.

Pavlas, D., (2010). A Model of Flow and Play in Game-based Learning: The Impact of Game Characteristics, Player Traits, and Player States. A Ph.D. dissertation in Applied Experimental and Human Factors Psychology, Department of Psychology, College of Sciences, University of Central Florida, Orlando, Florida, USA.

Pekrun, R. (1992). The Impact of Emotions on Learning and Achievement: Towards a Theory of Cognitive/Motivational Mediators. Journal of Applied Psychology, 41 (Oct. 1992), 359-376. doi= http://dx.doi.org/10.1111/j.1464-0597.1992.tb00712.x.

Psotka, J., & Mutter, S. A. (1988). Intelligent Tutoring Systems: Lessons Learned. Lawrence Erlbaum Associates. ISBN 0-8058-0192-8.

Rodriguez, H., Beck, D., Lind, D., & Lok, B. (2008). Audio Analysis of Human/Virtual-Human Interaction. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), 5208, 154-161. Springer, Heidelberg.

Saragih, J., Lucey, S. & Cohn, J. F. (2011). Deformable Model Fitting by Regularized Landmark Mean-Shift. International Journal of Computer Vision (IJCV), 91(2), 200-215.

Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., & Bigdeli, A. (2008). How Do You Know that I Don't Understand? A Look at the Future of Intelligent Tutoring Systems. Computers in Human Behavior, 24(4), 1342-1363. doi: 10.1016/j.chb.2007.07.008.

Sebe, N. (2009). Multimodal Interfaces: Challenges and Perspectives. Journal of Ambient Intelligence and Smart Environments, 1(1), 23-30.

Tijs, T., Brokken, D., & IJsselsteijn, W. (2009). Creating an Emotionally Adaptive Game. 7th International Conference in Entertainment Computing (ICEC 2008), 122-133. Springer Berlin Heidelberg, Pittsburgh, PA, USA.

Schwarz, N. (1990). Feeling as Information. Informational and Motivational Functions of Affective States. In E.T. Higgins & R. Sorrentino (Eds.). Handbook of Motivation and Cognition. Foundations of Social Behavior. New York: Guilford Press.

Van der Molen, H. T., & Gramsbergen-Hoogland, Y. H. (2005). Communication in Organizations: Basic Skills and Conversation models. New York, NY: Psychology Press.

Van Eck, R. (2010, page 115). Interdisciplinary Models and Tools for Serious Games: Emerging Concepts and Future Directions 1st Edition. Series: Premier Reference Source, Information Science Publishing. ISBN-10: 1615207198.

Vorvick, L., Avnon, T., Emmett, R., & Robins, L. (2008). Improving Teaching by Teaching Feedback. Med Educ, 42, 513-43.

Wallbott, H. G. (1998). Bodily Expression of Emotion. European Journal of Social Psychology, 28(6), 879-896.

Westera, W. (2013). The Digital Turn. How the Internet Transforms Our Existence. Bloomington, In: Author house.

Westera, W., Nadolski, R., Hummel, H. G. K., & Wopereis, I. (2008). Serious Games for Higher Education: A Framework for Reducing Design Complexity. Journal of Computer Assisted Learning, 24(5), 420-432.

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van der Spek, E. D. (2013). A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games. Journal of Educational Psychology, 105 (2), 249-265.

Wu, C.H., Huang, Y.M., & Hwang, J.P. (2015). Review of affective computing in education/learning: Trends and challenges. British Journal of Educational Technology. doi = 10.1111/bjet.12324.

Zhang, T., Hasegawa-Johnson, M., & Levinson, S. E. (2003). Mental State Detection of Dialogue System Users via Spoken Language. In ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR). Kyoto, Japan.